An efficient computer vision framework for RSU-based accident recognition and V2X communication

Danish Ather[1], Miraziz Talipov[2]
[1]Amity University, Noida, India
[1]Tashkent University of Information Technologies named after Muhammad al-Khwarizmi, Amir Temur Avenue 108, Tashkent, Uzbekistan
[2]Tashkent State Transport University, 1 Temiryulchilar Street, 100167, Tashkent, Uzbekistan
[1]Corresponding author
E-mail: [1]*danishather@gmail.com*, [2]*talipov_m@tstu.uz*

Received 30 January 2026; accepted 27 April 2026; published online 19 August 2026
DOI https://doi.org/10.21595/mme.2026.26075

Check for updates

Copyright © 2026 Danish Ather, et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract. Road traffic accidents require rapid detection and timely warning dissemination to reduce secondary collisions and improve emergency response. This paper presents an efficient roadside-unit-based computer vision framework for real-time accident recognition integrated with Vehicle-to-Everything communication. The proposed architecture combines lightweight convolutional spatial feature extraction, optical-flow-based motion analysis, temporal score aggregation, and low-latency V2X alert generation within an edge-deployable RSU pipeline. A formal system model is introduced to describe accident likelihood estimation, detection decision making, latency constraints, and communication reliability requirements. The experimental evaluation was conducted on a traffic video dataset containing 3,300 clips with normal traffic, single-vehicle accidents, multi-vehicle collisions, and non-accident anomalies. The proposed hybrid framework achieved 95.8 % accuracy, 95.0 % precision, 94.6 % recall, and 94.8 % F1-score, outperforming CNN-only and motion-only baselines. The average processing latency was approximately 40 ms per frame, indicating the feasibility of real-time operation on RSU-grade embedded hardware. These results show that the proposed framework can provide accurate and timely accident detection while supporting rapid V2X warning dissemination for next-generation intelligent transportation systems.

Keywords: roadside unit, traffic accident detection, computer vision, edge intelligence, V2X communication, intelligent transportation systems.

Nomenclature

t	Discrete time index for video frames
I_t	RGB frame captured by the RSU camera at time t
F_t	CNN feature vector extracted from frame I_t
M_t	Optical flow field between I_t and I_{t+1}
$\|\|M_t\|\|$	Magnitude map of optical flow at time t
$S_(t)$	Accident likelihood score at time t
θ	Decision threshold for accident declaration
λ	Traffic flow intensity (vehicles per unit time)
τ_{inf}	Inference time of the RSU vision module
τ_{tx}	V2X transmission delay for safety messages
τ_{tot}	Total detection-to-alert latency
P_d	Probability of correct accident detection
P_{fa}	Probability of false accident alarm
P_{cov}	Probability that a vehicle is within RSU coverage
P_{succ}	Probability of successful V2X message reception

1. Introduction

Road traffic accidents remain a major challenge for modern transport systems because delayed incident recognition can increase injury severity, traffic congestion, and the risk of secondary collisions. In conventional road surveillance practice, accident identification often depends on fixed cameras, manual monitoring, or delayed reports from road users, which reduces the speed and reliability of emergency response. For this reason, intelligent transportation systems are increasingly moving toward automated perception and cooperative safety mechanisms capable of detecting hazardous events and disseminating warnings in real time.

At the same time, Vehicle-to-Everything communication has created new opportunities for safety-oriented interaction between vehicles, roadside infrastructure, and traffic management services. Technologies such as Dedicated Short-Range Communications and cellular V2X enable roadside units to broadcast event notifications, support cooperative awareness, and improve local traffic intelligence [1-5]. In parallel, recent progress in edge intelligence and roadside infrastructure has shown that RSUs can evolve from passive communication relays into intelligent sensing and processing nodes capable of performing localized perception tasks with low latency [6-9].

Computer vision has become one of the most promising directions for automated traffic monitoring. Deep learning models have significantly improved vehicle detection, scene understanding, anomaly recognition, and traffic event analysis, while optical-flow-based methods remain useful for capturing abrupt motion changes and irregular scene dynamics [10-13]. However, many high-accuracy vision solutions are computationally demanding and are designed primarily for cloud-based processing or offline analysis. This limits their direct applicability to safety-critical roadside deployment, where the system must operate under constrained computational resources and strict delay requirements.

Another important research direction concerns cooperative perception and vehicle-infrastructure interaction. Recent studies have demonstrated the value of roadside sensing, edge processing, and V2X-assisted environmental awareness, especially at intersections and occluded road segments [7], [14-17]. In addition, large-scale cooperative perception datasets have improved the reproducibility of research on vehicle-infrastructure systems and enabled more realistic evaluation scenarios [18-22]. Nevertheless, the existing literature still tends to address accident recognition, roadside perception, and V2X message dissemination as partially separate problems rather than as a unified end-to-end safety pipeline.

Therefore, an important gap remains in the design of lightweight RSU-based frameworks that can simultaneously detect accidents from roadside video, operate under edge-computing constraints, and immediately generate safety alerts for nearby vehicles. This need is especially relevant for urban road segments, intersections, and critical corridors where rapid local detection and warning dissemination can substantially reduce response time and downstream risk.

The novelty of this study lies in three main aspects. First, the paper proposes an RSU-edge accident detection framework that combines convolutional spatial feature extraction and optical-flow-based motion analysis in a unified real-time roadside perception pipeline. Second, unlike approaches focused only on visual recognition or only on vehicular communication, the proposed method explicitly integrates accident perception with V2X safety-message dissemination. Third, the study evaluates the trade-off between detection performance and response latency for deployment-oriented roadside operation, which is essential for practical intelligent transportation systems [4-7]. More generally, the work follows the broader trend of applying computational and data-driven methods to transport engineering problems and safety-oriented system design [23-25].

The main contributions of the paper can be summarized as follows.

An integrated architecture is developed for RSU-based accident recognition, decision making, and V2X alert broadcasting.

A hybrid vision pipeline is introduced that combines lightweight convolutional feature

extraction with motion-consistency analysis to estimate accident likelihood in real time.

A formal system model is presented to describe accident scoring, threshold-based decision logic, latency components, and communication reliability constraints.

The proposed framework is experimentally evaluated from the perspective of both detection quality and end-to-end operational feasibility in roadside intelligent transportation scenarios.

The remainder of the paper is organized as follows. Section 2 reviews related work on vision-based traffic incident detection, roadside intelligence, and V2X communication. Section 3 presents the system model and problem formulation. Section 4 describes the proposed RSU-based vision architecture. Section 5 explains the accident recognition and decision logic. Section 6 presents the V2X communication design. Section 7 describes the experimental setup. Section 8 reports and discusses the results. Section 9 concludes the paper and outlines future research directions.

2. Background and related work

2.1. Computer vision for traffic incident detection

Computer vision has become one of the main technological directions for automated traffic monitoring, including vehicle detection, trajectory analysis, congestion estimation, anomaly recognition, and incident detection. Early traffic-scene analysis methods mainly relied on background subtraction, motion segmentation, handcrafted descriptors, and rule-based event interpretation. Although such approaches are computationally lightweight, they are often sensitive to illumination changes, shadows, camera vibrations, and scene complexity, which limits their robustness in real roadside environments [11], [18].

The development of deep learning has substantially improved visual scene understanding in transportation applications. Convolutional neural networks have been widely adopted for object detection, classification, and event recognition, while modern visual models can capture richer scene semantics than classical handcrafted approaches [10], [13]. In accident and near-miss analysis, several studies have also explored spatio-temporal architectures in which spatial perception is combined with temporal reasoning. However, such models may become computationally demanding when deployed on embedded roadside hardware, especially under strict latency requirements.

Motion information remains particularly important for accident recognition because collisions, abrupt stops, skidding, rollover-like behavior, and irregular trajectories are strongly reflected in short-term scene dynamics. Optical-flow-based representations provide dense motion fields between consecutive frames and therefore offer an explicit description of temporal changes in the traffic scene [11], [12]. For this reason, hybrid approaches that combine appearance-based spatial features with motion-based temporal descriptors are especially attractive for safety-critical accident detection, since they can improve sensitivity to abnormal traffic events while preserving interpretability of movement-related evidence [12], [13].

At the same time, recent surveys indicate that despite considerable progress in video-based accident detection, challenges remain in achieving robust real-time performance under weather variation, occlusion, scale changes, and deployment constraints typical of intelligent transportation environments [13]. These limitations are particularly relevant when the perception pipeline must run directly on a roadside unit rather than in a cloud environment.

2.2. RSUs, edge intelligence, and V2X communication

Vehicle-to-Everything communication has become a core component of intelligent transportation systems because it enables cooperative awareness, warning dissemination, and information exchange between vehicles, infrastructure, and traffic-management services. Foundational DSRC-based communication frameworks and later vehicular communication

studies established the basis for safety-oriented roadside messaging, low-latency broadcasting, and infrastructure-assisted road monitoring [1-3]. More recent V2X developments, including 5G NR V2X and standardized decentralized environmental notification services, further strengthen the communication support for time-critical road-hazard alerts [4], [5].

Within such systems, roadside units are no longer viewed only as fixed communication relays. Instead, they are increasingly treated as intelligent edge nodes capable of integrating local sensing, message generation, and distributed decision support. Recent surveys show that roadside infrastructure can provide persistent scene coverage, stable sensing geometry, and localized intelligence, making it particularly suitable for applications at intersections, highway segments, ramps, and other safety-critical road locations [6-9]. In contrast to onboard sensing alone, RSU-based perception can better observe occluded or conflict-prone zones and can assist multiple nearby vehicles simultaneously.

Another major research direction is cooperative vehicle-infrastructure perception. Recent studies have demonstrated the value of roadside sensing systems, intersection-level cooperative vision, and field-tested roadside perception pipelines for improving traffic awareness and supporting automated driving functions [14-17]. In parallel, the emergence of large-scale vehicle-infrastructure datasets has significantly improved reproducibility in this area. Datasets such as KITTI and more recent cooperative perception benchmarks, including DAIR-V2X, V2X-Seq, TUMTraf V2X, and V2X-Real, provide increasingly realistic foundations for evaluating traffic-scene perception and infrastructure-assisted intelligence [18-22].

Despite these advances, the deployment of computer vision directly on RSUs still introduces practical constraints related to processing power, memory consumption, communication deadlines, and system scalability. For safety-critical applications, cloud-only processing may introduce undesirable delay, while full reliance on vehicle-side sensing may limit spatial awareness. Consequently, edge-based RSU perception combined with direct V2X warning dissemination is increasingly regarded as a promising architecture for low-latency intelligent transportation services [4], [6], [7].

2.3. Research gap and positioning of the present study

The existing literature demonstrates strong progress in three partially overlapping areas: vision-based traffic accident detection, roadside and cooperative perception, and V2X safety communication. However, these directions are often studied separately. Many works focus primarily on visual recognition quality without explicitly addressing roadside deployment constraints, while communication-oriented studies often emphasize protocol behavior rather than the perception process that triggers safety alerts [4], [6], [7], [13], [14].

In addition, although roadside intelligence is increasingly recognized as an important ITS component, fewer studies present a unified RSU-centered pipeline that combines lightweight computer vision, motion-aware accident scoring, latency-conscious edge execution, and direct V2X warning generation within a single framework. This leaves several practical gaps: limited attention to lightweight accident-detection pipelines optimized for RSU hardware, insufficient coupling between perception outputs and warning-message generation, and incomplete end-to-end consideration of the time interval between accident occurrence, detection, and dissemination of driver alerts [4-7].

The present study is positioned in this gap. It proposes an integrated RSU-based framework that combines convolutional spatial feature extraction, optical-flow-based motion analysis, threshold-based accident decision logic, and V2X warning dissemination within an edge-deployable architecture. In this way, the work aims to connect visual accident recognition with communication-aware roadside operation and to evaluate the trade-off between detection performance and real-time responsiveness in practical intelligent transportation scenarios.

3. System model and problem formulation

3.1. Traffic scene and RSU deployment

Consider a road segment monitored by a single roadside unit equipped with a fixed camera and a local embedded processing module. The camera covers one or several traffic lanes within a predefined field of view and continuously captures video frames at a rate of f_{cam} frames per second. Each frame is transmitted to the RSU processing unit, where the traffic scene is analyzed in streaming mode for accident recognition.

Let $I_t \in R^{H\times W\times 3}$ denote the RGB image captured at time t, where H and W are the frame height and width, respectively. Based on the incoming video stream and recent temporal evidence, the RSU computes an accident likelihood score $S_t \in [0,1]$ for each time instant. This score reflects the confidence of the vision module that an accident or accident-like abnormal event is occurring in the observed traffic scene.

The considered deployment assumes roadside sensing at the network edge, where perception, decision making, and warning generation are performed locally. This assumption is important because safety-oriented accident detection requires low response latency and should not depend exclusively on remote cloud processing.

3.2. Accident state and detection decision

The true accident condition of the monitored scene is represented by a binary variable $A_t \in \{0,1\}$:

$$A_t = \begin{cases} 1, & \text{if an accident is occurring in the scene at time } t, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Eq. (1) defines the hidden ground-truth state of the traffic scene. In practice, this state is not directly observable by the RSU and must be inferred from visual evidence extracted from the video stream.

Instead of observing A_t directly, the RSU estimates an accident likelihood score S_t using the vision pipeline. A binary decision A_t^* is then obtained by thresholding this score:

$$A_t^* = \begin{cases} 1, & S_t \geq \theta, \\ 0, & S_t < \theta, \end{cases} \tag{2}$$

where θ is a configurable decision threshold.

Eq. (2) converts the continuous accident likelihood into an operational detection output used by the warning subsystem. The threshold θ controls the trade-off between sensitivity and false alarms and is later selected on a validation set according to the required safety level.

The design objective is to construct the mapping from the observed frames $\{I_t\}$ to the scores $\{S_t\}$ in such a way that the probability of detection: $P_d = P_r\,(A_t^* = 1|A_t = 1)$, is maximized while the probability $P_{fa} = P_r(A_t^* = 1 \mid A_t = 0)$ remains sufficiently low for practical roadside deployment.

3.3. Latency and reliability constraints

Once an accident is detected, the RSU generates and broadcasts a safety alert to nearby vehicles through the V2X communication subsystem. For such alerts to be useful in practice, the total delay from accident occurrence to message delivery must remain bounded. This requirement can be expressed as:

$$\tau_{tot} = \tau_{det} + \tau_{tx} \leq \tau_{max}, \tag{3}$$

where τ_{det} is the detection delay produced by perception and decision making, τ_{tx} is the communication delay, and τ_{max} is the maximum acceptable response time for the target safety application.

Eq. (3) formalizes the real-time requirement of the proposed framework. It shows that high recognition accuracy alone is not sufficient; the accident detection pipeline must also satisfy strict end-to-end timing constraints in order to support timely driver warning and emergency response.

In addition to latency, the warning mechanism must satisfy basic reliability conditions. Let P_{succ} denote the probability that a vehicle within the target area successfully receives the accident alert and let P_{cov} denote the probability that a relevant vehicle lies inside the RSU coverage region. Then the system should satisfy:

$$P_{succ} \geq P_{succ,min}, \quad P_{cov} \geq P_{cov,min}, \tag{4}$$

where $P_{succ,min}$ and $P_{cov,min}$ are application-dependent minimum targets.

Eq. (4) emphasizes that the practical value of an RSU-based accident warning system depends not only on correct visual detection but also on adequate communication reach and successful message reception.

3.4. Problem statement

Given the RSU hardware characteristics, the camera configuration, and the V2X communication stack, the problem addressed in this study is to design:

a) a vision-based mapping $g; (I_t, I_{t-1}, \dots) \mapsto S_t$, which estimates the accident likelihood from spatial and temporal visual evidence.

b) a decision rule parameterized by the threshold θ for converting the continuous score into an accident declaration.

c) a low-latency roadside processing pipeline capable of satisfying the timing condition in (3).

d) a communication-aware alert generation mechanism that supports the reliability conditions in (4).

Accordingly, the proposed framework seeks to balance four practical requirements: detection accuracy, robustness to abnormal motion patterns, low inference latency, and timely dissemination of roadside safety alerts. The following sections describe how this objective is addressed through the hybrid vision architecture, the accident recognition algorithm, and the V2X messaging design.

4. Proposed RSU-based computer vision framework

4.1. Overall architecture

The proposed RSU-based framework is designed as a lightweight roadside perception pipeline for real-time accident recognition and immediate V2X warning generation. The overall processing chain consists of four main stages: 1) frame pre-processing and region-of-interest selection, 2) CNN-based spatial feature extraction, 3) optical-flow-based motion analysis, and 4) feature fusion, accident scoring, and threshold-based decision making. Once an accident is detected, the RSU triggers the V2X alerting module, which generates and broadcasts a roadside warning message to nearby vehicles and traffic-management entities.

The architecture is intentionally organized around edge deployment constraints. In contrast to cloud-centered processing, the proposed design performs scene interpretation locally at the RSU in order to reduce response delay and support safety-critical operation under limited computation and communication budgets. The combination of spatial appearance cues and motion-consistency cues is intended to improve recognition robustness in complex roadside scenes, where accidents may involve abrupt deceleration, unusual trajectories, lateral deviations, or stationary post-impact configurations. The overall architecture of the proposed RSU-based accident recognition and V2X

warning framework is shown in Fig. 1.

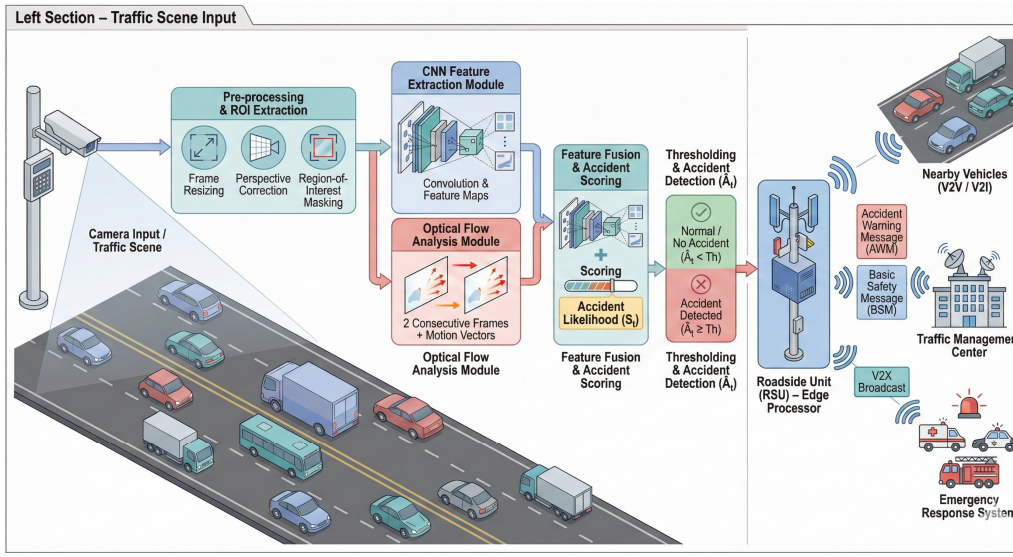


Fig. 1. Overall architecture of the proposed RSU-based accident recognition and V2X communication framework. The figure is an original schematic diagram created and designed by the authors specifically for this manuscript; no third-party copyrighted material was reused

4.2. Pre-processing and region-of-interest selection

Each incoming frame I_t is first resized to a fixed resolution $H' \times W'$ (specifically, 640×480 pixels in our experimental setup) and normalized to ensure stable inference under varying roadside image conditions. Depending on the camera installation geometry, an optional perspective transformation may be applied to obtain a more regular road-plane representation, which facilitates motion interpretation and reduces geometric distortion.

To suppress irrelevant background content, a region-of-interest mask R is applied so that the subsequent feature extraction stages focus primarily on roadway, lane, and shoulder regions. The masked pre-processed frame is written as:

$$I_t^{*(R)}(x,y) = \begin{cases} I_t^*(x,y), & (x,y) \in R, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

Eq. (5) formalizes the ROI filtering step, which reduces interference from buildings, sky regions, sidewalks, and other non-traffic objects. This step is particularly useful for roadside cameras with a wide field of view, where non-road pixels may otherwise introduce unnecessary variation into both spatial and motion descriptors.

4.3. CNN-based spatial feature extraction

After pre-processing, the masked frame $I_t^{*(R)}$ is passed through a lightweight convolutional neural network to extract compact spatial descriptors:

$$F_t = \phi_{CNN}\left(I_t^{*(R)}; \Theta\right), \tag{6}$$

where Θ denotes the trainable parameters of the CNN and $F_t \in R^d$ is the resulting feature vector.

Eq. (6) maps the visual content of the current frame into a low-dimensional representation that

encodes appearance-based accident evidence such as abnormal vehicle posture, overlap between vehicles, debris-like structures, stopped vehicles in unusual positions, and other scene irregularities. The CNN is intentionally kept lightweight in order to support low-latency execution on RSU-grade embedded hardware. Depthwise separable convolutions, reduced channel width, or other compact architectural strategies may be used to preserve an acceptable balance between representation quality and inference speed.

4.4. Optical-flow-based motion analysis

Although spatial appearance is important, accident recognition also depends strongly on short-term motion behavior. Therefore, the framework computes dense optical flow between two consecutive pre-processed frames:

$$M_t = \psi\left(I_t^{*(R)}, I_{t-1}^{*(R)}\right), \tag{7}$$

where $\psi(\cdot)$ denotes the optical flow operator and M_t is the resulting motion field.

Each pixel is associated with a 2D motion vector $\left(u_t(x,y), v_t(x,y)\right)$, and the corresponding motion magnitude map is defined as:

$$\|M_t\|(x,y) = \sqrt{u_t(x,y)^2 + v_t(x,y)^2} \tag{8}$$

Eq. (8) provides a scalar motion intensity representation that highlights abrupt displacement patterns, irregular deceleration, skidding, collision-related impact motion, and sudden directional deviations. To obtain compact motion descriptors suitable for low-cost accident scoring, summary statistics are computed over the ROI:

$$\mu_t = \frac{1}{|R|}\sum_{(x,y)\in R} \|M_t\|(x,y), \tag{9}$$

$$\sigma_t^2 = \frac{1}{|R|}\sum_{(x,y)\in R} (\|M_t\|(x,y) - \mu_t)^2. \tag{10}$$

Eq. (9) captures the average motion magnitude within the monitored roadway region, while Eq. (10) reflects motion variability. In accident situations, both the mean motion level and its dispersion may change sharply, making these statistics useful indicators of abnormal traffic behavior. More detailed descriptors such as directional histograms or lane-wise sector statistics may also be incorporated if needed.

4.5. Feature fusion and accident scoring

To combine spatial and temporal evidence, the CNN feature vector and the motion descriptors are concatenated into a joint representation:

$$Z_t = \lceil F_t^T, \mu_t, \sigma_t^2 \rceil^T. \tag{11}$$

This fused vector is then mapped to an accident likelihood score through a compact fully connected decision layer:

$$S_t = \sigma(w^T Z_t + b), \tag{12}$$

where w and b are learnable parameters and $\sigma(\cdot)$ is the sigmoid function.

Eq. (12) transforms the fused spatial-motion representation into a frame-level estimate of accident probability. The fusion stage is important because some accident scenarios are dominated

by appearance cues, while others are more clearly expressed through motion irregularities. Joint modeling therefore improves robustness relative to single-modality baselines.

To reduce sensitivity to isolated noisy frames, short-term temporal averaging is applied:

$$S_t^* = \frac{1}{K}\sum_{k=0}^{K-1} S_{t-k}. \tag{13}$$

Eq. (13) smooths the frame-wise accident score over a sliding window of length K, thereby suppressing short spurious peaks caused by non-accident but abrupt traffic events. The decision rule introduced in Section 3 is then applied to S_t^* rather than to the raw score S_t, yielding a more stable accident declaration suitable for safety-oriented roadside operation.

5. Accident recognition algorithm

5.1. Frame-level accident scoring

Based on the fused representation Z_t defined in Section 4, the RSU computes a frame-level accident likelihood score S_t, which reflects the probability that the observed traffic scene contains an accident or a strong accident-like anomaly. In the proposed framework, this score is obtained from a compact fusion layer combining spatial and motion evidence, as expressed in Eq. (12). However, a single frame is often insufficient for reliable accident recognition, since abrupt but non-critical traffic maneuvers may temporarily resemble hazardous events.

For this reason, the frame-level score is interpreted as an instantaneous confidence measure rather than as a final alarm condition. The recognition module therefore uses the sequence of recent scores $\{S_t\}$ to estimate whether the observed event is persistent enough to justify an accident declaration. This design reduces the sensitivity of the system to short-duration noise, camera jitter, and isolated motion peaks caused by non-accident traffic disturbances.

In practice, an accident candidate emerges when the fused score increases beyond the normal operating range and remains consistently elevated during a short temporal interval. The objective of the decision algorithm is thus not only to maximize sensitivity to true collision events, but also to suppress false activations caused by braking waves, dense traffic compression, or temporary occlusions.

5.2. Temporal stabilization and event confirmation

To improve robustness, the proposed system applies sliding-window temporal averaging to the raw frame-level scores, as already defined in Eq. (13). The resulting stabilized score S_t^* is then compared with the decision threshold θ. A preliminary accident flag is produced as:

$$D_t = \begin{cases} 1, & S_t^* \geq \theta, \\ 0, & \text{otherwise.} \end{cases} \tag{14}$$

Eq. (14) defines the first-stage decision output of the recognition module. Its function is to identify time instants at which the smoothed accident confidence becomes sufficiently high to indicate a potentially critical event.

To avoid triggering an alert from a single unstable peak, the framework introduces a persistence requirement. Let L denote the minimum number of consecutive frames for which the preliminary flag remains active. Then the final confirmed accident declaration $\widehat{A_t}$ is written as:

$$\widehat{A_t} = \begin{cases} 1, & \sum_{i=0}^{L-1} D_{t-1} = L, \\ 0, & \text{otherwise.} \end{cases} \tag{15}$$

Eq. (15) acts as a short-term confirmation rule. It ensures that the system responds to sustained evidence of an accident rather than to isolated threshold crossings. This is especially important in roadside environments, where transient lighting changes, shadows, or abrupt but safe vehicle maneuvers may temporarily affect the raw decision score.

The parameters K in Eq. (13) and L in Eq. (15) jointly determine the temporal sensitivity of the detector. A smaller K or L generally reduces detection delay but may increase false alarms, whereas larger values improve stability at the cost of slower response. Therefore, these parameters must be selected as a compromise between responsiveness and robustness.

5.3. Threshold selection and operating point

The choice of threshold θ is critical because it governs the trade-off between missed detections and false alarms. In the proposed framework, the threshold is selected on a validation subset by analyzing the receiver operating characteristic and identifying an operating point that provides high recall while preserving acceptable precision for roadside deployment. The ROC curve used for selecting the operating threshold is shown in Fig. 2.

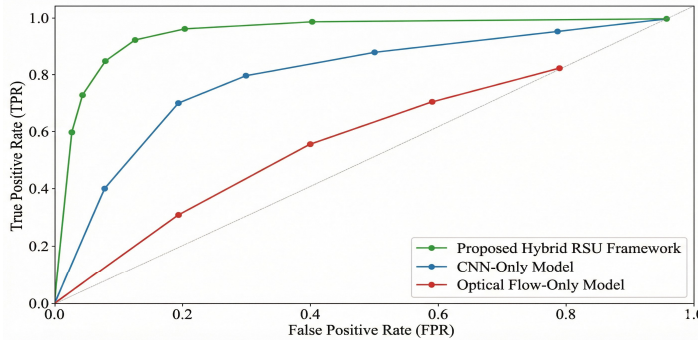


Fig. 2. ROC curve of the proposed accident detection framework for threshold selection

Based on this analysis, the selected threshold should satisfy two practical objectives. First, it should maintain high sensitivity to true accident events, since missed detections directly reduce the safety value of the system. Second, it should keep the false-alarm rate sufficiently low, because excessive warning messages may reduce driver trust and overload the communication channel in dense traffic conditions.

For completeness, the operating threshold may be selected using the following criterion:

$$\theta^* = argmax_\theta J(\theta), \tag{16}$$

where $J(\theta)$ is a scalar selection objective. In practical evaluation, $J(\theta)$ may correspond to the F1-score, Youden's index, balanced accuracy, or another deployment-oriented criterion depending on the relative importance of recall and false-alarm suppression.

If the F1-score is used as the primary validation criterion, then:

$$F1(\theta) = \frac{2P(\theta)R(\theta)}{P(\theta) + R(\theta)}, \tag{17}$$

where $P(\theta)$ and $R(\theta)$ are precision and recall obtained at threshold θ. Eq. (17) is useful when the objective is to balance sensitivity and precision under class imbalance or event rarity.

Additional sensitivity results for different temporal averaging windows are reported in Appendix Table A1. This supplementary analysis helps justify the selected value of K and demonstrates the effect of temporal stabilization on the final detection behavior.

5.4. Alarm generation logic

Once the confirmed accident state $\widehat{A_t} = 1$ is obtained, the RSU generates an internal event record containing the estimated event time, the monitored road segment identifier, and the accident confidence level. Let γ_t denote the severity-related confidence label derived from the stabilized score. A simple confidence mapping may be defined as:

$$\gamma_t = \begin{cases} \text{low}, & 0 \leq S_t^* < \theta_1, \\ \text{medium}, & \theta_1 \leq S_t^* < \theta_2, \\ \text{high}, & S_t^* \geq \theta_2, \end{cases} \tag{18}$$

where θ_1 and θ_2 are optional confidence thresholds satisfying $\theta \leq \theta_1 < \theta_2 \leq 1$.

Eq. (18) is not intended to estimate true crash severity in the forensic sense; rather, it provides an operational confidence label that can be used to prioritize roadside warning dissemination and traffic-management response. This makes the alerting mechanism more flexible in practical ITS deployment.

After confirmation, the accident event is passed to the communication layer, which encapsulates the event information into a V2X warning message and broadcasts it to nearby vehicles. The structure of this message and the associated dissemination logic are described in the next section.

6. V2X communication and alert dissemination

6.1. Alert message structure

Once the accident state has been confirmed by the recognition module, the RSU generates a structured roadside warning message for dissemination to nearby vehicles and, if required, to the traffic-management backend. The purpose of this message is to provide timely and compact information about the detected hazardous event so that surrounding vehicles can adapt their speed, lane choice, or route in advance. In the proposed framework, the warning packet is constructed directly at the RSU in order to minimize the delay between perception and communication.

The generated message contains the following core fields:

1) Message identifier and message type.
2) RSU identifier.
3) Detection timestamp.
4) Geographical location of the event.
5) Estimated accident confidence or severity level.
6) Lane and direction information, when available.
7) Optional contextual descriptors such as traffic density, weather condition, or road segment identifier.

These fields are sufficient to support a compact hazard-notification logic in which vehicles receive not only the existence of an accident alert, but also the minimum contextual information required for local driving response and traffic-management interpretation.

The structure of the generated roadside accident warning message is illustrated in Fig. 3.

To formalize the message payload, let the accident warning packet at time t be represented as:

$$m_t = \{ID_t, RSU_t, T_t, L_t, \gamma_t, \delta_t, C_t\}, \tag{19}$$

where ID_t is the message identifier, RSU_t is the roadside unit identifier, T_t is the detection time, L_t is the event location, γ_t is the confidence or severity label, δ_t denotes directional or lane-related metadata, and C_t represents optional contextual information.

Eq. (19) defines the communication-level abstraction of the detected accident event. Its

function is to map the perception output into a transportable warning object that can be broadcast through the V2X interface without exposing unnecessary raw sensing data.

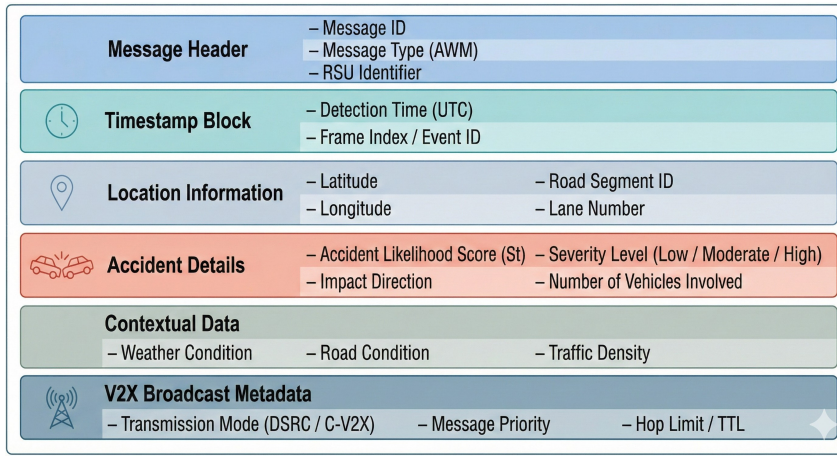


Fig. 3. Structure of the V2X accident warning message generated by the proposed RSU framework

6.2. Latency model

For accident warning systems, communication value depends not only on correct recognition but also on timely delivery. The total end-to-end delay from accident detection to message availability at the V2X interface can be modeled as:

$$\tau_{tot} = \tau_{inf} + \tau_{agg} + \tau_{tx}\,, \tag{20}$$

where τ_{inf} is the inference time of the perception module, τ_{agg} is the temporal aggregation and confirmation delay, and τ_{tx} is the transmission-related delay associated with message construction, channel access, and wireless delivery.

Eq. (20) is important because it explicitly links perception and communication into a single operational measure. Even a highly accurate detector becomes less useful if the alert is delayed beyond the time window in which nearby vehicles can still react safely.

The transmission component τ_{tx} depends on several factors, including channel quality, network load, medium-access conditions, prioritization policy, and the specific communication stack used by the RSU, such as DSRC or C-V2X. For safety-critical alerting, priority queuing and message preemption mechanisms should be used whenever supported by the communication subsystem.

In practical deployment, the system should satisfy $\tau_{tot} \leq \tau_{max}$, where τ_{max} is the maximum acceptable accident-to-alert delay for the target intelligent transportation application.

6.3. Coverage and successful message reception

An accident warning is useful only if the target vehicle is both located inside the effective RSU communication region and able to receive the transmitted message successfully. Following the logic already present in the draft, the alert success probability can be expressed as:

$$P_{succ} = P_{cov} \cdot P_{rx|cov}, \tag{21}$$

where P_{cov} is the probability that a relevant vehicle lies inside the RSU coverage area and $P_{rx|cov}$ is the conditional probability of successful reception given that the vehicle is in coverage.

Eq. (21) emphasizes that communication performance depends on both spatial availability and link reliability. This is especially relevant in urban intersections, dense corridors, and partially obstructed road segments, where geometric coverage and packet reception quality may vary significantly over time.

To improve $P_{rx|cov}$, the framework may employ repeated transmission, short message redundancy, or multi-hop forwarding through neighboring infrastructure or vehicles. Such mechanisms are particularly useful under channel congestion or partial obstruction conditions. The choice of redundancy level should, however, remain balanced against channel occupancy and message-priority constraints.

6.4. Dissemination logic and system integration

Once the accident event has been encoded into the message structure defined in Eq. (19), the RSU inserts the warning packet into the V2X transmission queue with safety-critical priority. The dissemination logic may support both local broadcast to nearby vehicles and forwarding to external traffic-management systems, depending on the deployed network architecture.

At the system level, the V2X module acts as the communication bridge between the roadside perception engine and the cooperative safety ecosystem. In this way, the proposed framework links visual accident recognition with communication-aware roadside operation and enables an integrated accident-to-alert pipeline suitable for intelligent transportation scenarios.

7. Implementation and experimental setup

7.1. Dataset and scenarios

The experimental evaluation was conducted using a traffic video dataset composed of realistic roadway scenes together with accident-oriented event instances. The dataset includes four scenario categories: normal traffic flow, single-vehicle accident events, multi-vehicle collisions, and non-accident anomalies such as stopped vehicles or pedestrian interference. The inclusion of non-accident anomalies is important because these events may visually resemble hazardous situations and therefore provide a more realistic basis for false-alarm analysis.

A summary of the dataset composition is presented in Table 1.

Table 1. Composition of the traffic video dataset used for RSU-based accident detection

Scenario type	Number of clips	Average duration (s)	Total duration (min)
Normal flow (no accident)	2 000	20	667
Single-vehicle accident	600	15	150
Multi-vehicle collision	400	18	120
Non-accident anomalies	300	15	75
Total	3 300	–	1 012

The dataset contains a total of 3,300 clips, including 2,000 normal-flow clips, 600 single-vehicle accident clips, 400 multi-vehicle collision clips, and 300 non-accident anomaly clips. The total duration is approximately 1,012 min. The dataset was divided into training, validation, and test subsets using a stratified split in order to preserve the class distribution across all scenario types. To avoid temporal leakage, clips from the same event sequence were not distributed across different subsets.

7.2. RSU hardware and software environment

The proposed framework was implemented on an RSU-oriented embedded computing platform equipped with an ARM-based processor, memory-constrained edge resources, and a V2X communication interface. The prototype used a quad-core ARM processor (representing

edge-grade boards such as Raspberry Pi 4B or NVIDIA Jetson Nano), 4 GB RAM, and hardware support for floating-point operations. The software environment was based on a lightweight deep-learning library optimized for edge devices, while optical-flow computation was performed using an efficient real-time pipeline configured for roadside execution. The communication layer was designed to support DSRC or C-V2X message dissemination.

The hardware and software configuration of the RSU prototype used in the experiments is summarized in Table 2.

Table 2. RSU hardware and software configuration used in the experiments

Item	Specification
Computing platform	Embedded RSU-oriented edge-computing platform
Processor	Quad-core ARM processor (e.g., Broadcom BCM2711 or equivalent edge CPU)
Hardware acceleration	Graphics acceleration and floating-point support
Memory	4-8 GB RAM
Vision framework	Lightweight deep-learning library optimized for edge devices
CNN model type	Lightweight convolutional neural network for spatial feature extraction
Optical flow module	Efficient real-time optical-flow algorithm
ROI processing	Frame pre-processing, normalization, and region-of-interest selection
V2X interface	DSRC or C-V2X communication module
Message generation	On-device RSU alert generation and broadcast
Deployment mode	Local edge inference on the RSU without mandatory cloud offloading

This hardware-software configuration reflects the intended deployment scenario of the proposed framework, in which accident recognition and warning generation are performed locally at the roadside unit in order to minimize response delay and reduce dependence on remote cloud processing.

7.3. Training and evaluation protocol

The CNN feature extractor and the accident scoring module were trained using labeled accident and non-accident samples. The training objective was binary accident classification, and the loss function was defined as:

$$L = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log S_i + (1-y_i)\log(1-S_i)], \tag{22}$$

where $y_i \in \{0,1\}$ is the ground-truth label for sample i, S_i is the predicted accident likelihood, and N is the batch size.

The operating threshold θ was selected on the validation subset using ROC-based operating-point analysis. Performance was then evaluated on the held-out test subset using accuracy, precision, recall, and F1-score for accident recognition, together with latency analysis for real-time deployment assessment. In addition, confusion-matrix analysis was used to visualize class-wise prediction behavior, while latency profiling was performed separately for pre-processing, CNN inference, optical-flow computation, accident scoring, and V2X message generation.

The training and evaluation settings of the proposed RSU accident detection framework are summarized in Table 3.

The current experimental protocol defines the overall training-validation-test workflow, the binary accident classification objective, and the evaluation metrics used on the held-out test subset. This configuration provides a consistent basis for assessing accident-recognition quality together with real-time roadside processing capability.

Table 3. Training and evaluation settings of the proposed RSU accident detection framework

Item	Setting
Dataset partition	Training / validation / test split
Training target	Joint training of CNN feature extractor and accident scoring module
Training labels	Accident / non-accident labels
Loss function	Binary cross-entropy
Threshold selection	Validation set with ROC-based operating-point selection
Evaluation subset	Test set
Reported metrics	Accuracy, precision, recall, F1-score, latency
Confusion-matrix analysis	Yes
Event categories used in evaluation	Normal flow, single-vehicle accident, multi-vehicle collision, non-accident anomalies
Total number of clips	3,300
Total dataset duration	1,012 min

8. Results and discussion

8.1. Detection performance

The proposed RSU-based framework demonstrated strong accident-recognition performance across diverse traffic-event categories, including normal flow, single-vehicle accidents, multi-vehicle collisions, and non-accident anomalies. The quantitative results are summarized in Table 4, where the proposed hybrid approach is compared with two reference baselines: a CNN-only method based on spatial appearance cues and a motion-only method based on optical-flow dynamics. The current manuscript reports 95.8 % accuracy, 95.0 % precision, 94.6 % recall, and 94.8 % F1-score for the proposed hybrid framework, which is higher than both single-modality baselines.

Table 4. Accident detection performance of different methods

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
CNN-only (spatial features)	90.4	89.2	88.7	88.9
Motion-only (optical flow)	87.1	85.0	86.5	85.7
Proposed hybrid RSU framework	95.8	95.0	94.6	94.8

The results indicate that combining spatial and motion information provides a clear advantage over single-stream processing. The CNN-only baseline captures appearance-level evidence such as damaged vehicle configurations, overlap patterns, and abnormal spatial layouts, whereas the motion-only baseline is more sensitive to abrupt trajectory changes and short-term dynamic irregularities. However, each of these streams alone remains incomplete. Their fusion improves both sensitivity to collision-like events and resistance to false positives caused by non-accident anomalies.

In practical terms, the observed performance gain suggests that accident recognition in roadside environments benefits from explicitly modeling both visual appearance and scene dynamics. This is especially important for RSU deployment, where the same camera must distinguish between true hazardous events and visually similar but operationally harmless conditions such as temporary stops, traffic compression, or pedestrian interference. The class-wise prediction behavior of the proposed framework is further illustrated by the confusion matrix in Fig. 4.

The confusion matrix in Fig. 4 helps visualize the remaining misclassification patterns. Although the proposed approach substantially reduces confusion between accident and non-accident conditions, certain difficult cases may still arise when abrupt but safe maneuvers produce accident-like motion signatures or when partial occlusion weakens the visibility of the impact event. Nevertheless, the overall confusion pattern remains consistent with the quantitative superiority of the hybrid framework shown in Table 4.

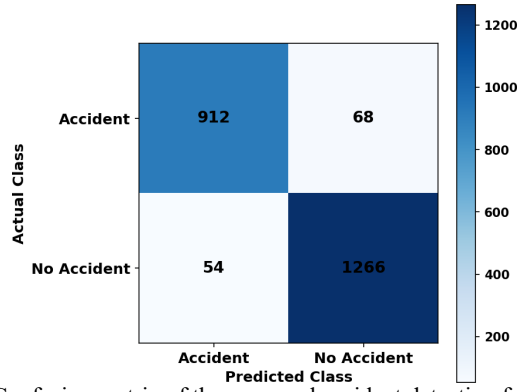


Fig. 4. Confusion matrix of the proposed accident detection framework

The comparative performance of the CNN-only, motion-only, and proposed hybrid RSU framework is further illustrated in Fig. 5.

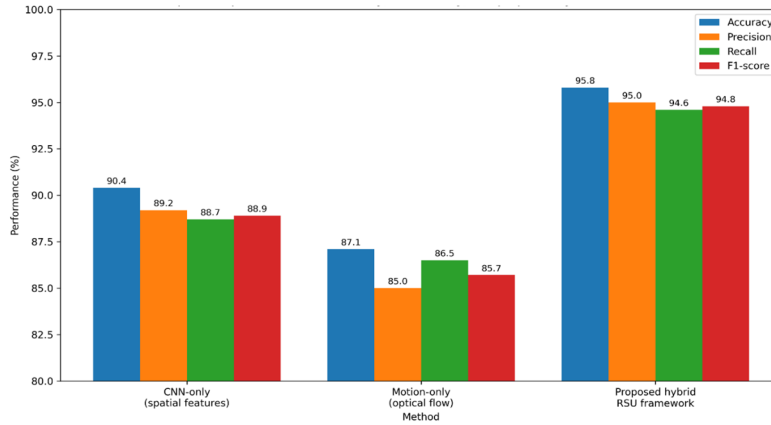


Fig. 5. Comparative performance of CNN-only, motion-only, and proposed hybrid RSU framework in terms of accuracy, precision, recall, and F1-score

8.2. Latency analysis

In addition to recognition quality, the practical suitability of the framework depends on its ability to satisfy real-time roadside constraints. The current experimental results show that the average end-to-end processing time is approximately 40 ms per frame on the RSU prototype, with CNN feature extraction and optical-flow computation being the dominant contributors to latency. The detailed breakdown is presented in Table 5.

Table 5. Latency breakdown of the proposed RSU accident detection pipeline

Component	Average time per frame (ms)	Relative share (%)
Pre-processing	4	10
CNN feature extraction	16	40
Optical flow computation	14	35
Accident scoring	3	7
V2X message construction	3	8
Total	40	100

According to the current measurements, pre-processing requires 4 ms, CNN feature extraction 16 ms, optical-flow computation 14 ms, accident scoring 3 ms, and V2X message construction

3 ms, giving a total of 40 ms per frame. These results indicate that the proposed architecture is feasible for real-time edge deployment under the considered RSU configuration. The latency profile also shows that further optimization efforts should primarily focus on the CNN inference stage and the motion-analysis branch, since together they account for the largest share of the total processing budget. The temporal sequence from accident occurrence to roadside detection and vehicle warning dissemination is illustrated in Fig. 6.

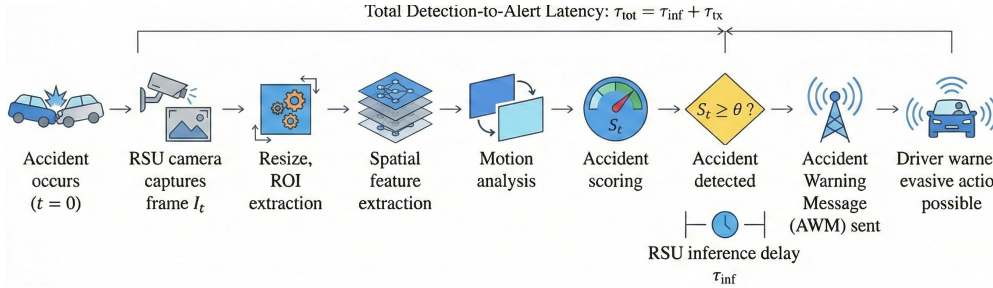


Fig. 6. Timeline from accident occurrence to roadside detection and V2X alert dissemination. The figure is an original schematic diagram created and designed by the authors specifically for this manuscript; no third-party copyrighted material was reused

Fig. 6 provides a process-level view of the accident-to-alert sequence. It illustrates that the practical value of the framework does not depend only on accurate classification, but also on the timely transition from visual perception to warning generation and message dissemination.

8.3. Robustness to environmental conditions

Preliminary observations indicate that the proposed framework maintains reasonable performance under moderate variations in illumination, camera angle, and common traffic-scene variability. This suggests that the hybrid combination of spatial and motion cues offers a degree of robustness to ordinary roadside environmental changes. However, the current study also indicates that performance may degrade under severe weather conditions such as heavy rain, dense fog, or snowfall, which reduce scene visibility and distort optical-flow estimation.

From a deployment perspective, this limitation is important because safety-critical systems must remain functional under non-ideal operating conditions. Therefore, one of the most promising directions for future research is the integration of additional sensing modalities, such as radar, thermal imaging, or infrastructure-level sensor fusion, in order to improve reliability in visually degraded environments.

8.4. Discussion on deployment and comparison with existing directions

The results support the practical relevance of RSUs as edge-intelligence nodes for safety-oriented transport monitoring. Compared with vehicle-only sensing, roadside deployment offers a stable observation geometry and can simultaneously support multiple approaching vehicles. Compared with cloud-centered accident analysis, local RSU inference reduces dependence on remote communication and better aligns with low-latency safety requirements. These advantages make RSU-based accident detection particularly attractive for intersections, highway segments, ramps, and other conflict-prone road locations.

At the same time, successful real-world deployment requires more than perception quality alone. It also depends on RSU placement strategy, communication coverage, backhaul availability, and integration with traffic-management infrastructure. In this sense, the proposed study should be viewed as a deployment-oriented framework rather than a fully generalized final solution. Its main contribution lies in showing that an integrated accident-recognition and

V2X-warning pipeline can be implemented under realistic edge constraints while preserving both strong recognition quality and acceptable processing delay.

Furthermore, while direct numerical comparison across different datasets must be interpreted cautiously, the achieved accuracy of 95.8 % aligns competitively with recent state-of-the-art vision-based accident detection methods. While cloud-heavy deep learning models often report accuracies in the 92-96 % range, achieving 95.8 % locally on an edge device highlights the efficiency of the proposed spatial-temporal fusion. A qualitative comparison with related studies is provided in Table 6. In contrast to prior works that often focus on only one or two components of the problem, the proposed framework combines vision-based accident recognition, RSU-centered deployment, edge-oriented processing, V2X alert generation, and explicit latency-aware operation in a unified pipeline.

Table 6. Qualitative comparison of the proposed framework with related studies

Study	Vision-based accident / perception	Roadside / RSU deployment	Edge-oriented execution	V2X alerting integration	Latency discussion
Fang et al. [13]	Yes	No	Limited	No	Limited
Creß et al. [7]	Yes	Yes	Yes	Limited	Limited
Huang et al. [14]	Yes	Yes	Yes	Yes	Limited
Masi et al. [15]	Yes	Yes	Limited	Limited	No
Mo et al. [17]	Yes	Yes	Yes	Yes	Limited
Proposed framework	Yes	Yes	Yes	Yes	Yes

9. Conclusions

This paper presented an RSU-based computer vision framework for real-time accident recognition integrated with V2X communication. The proposed system combines lightweight CNN-based spatial feature extraction, optical-flow-based motion analysis, temporal score stabilization, and roadside alert-message generation within a unified edge-deployable architecture. In this way, the study addressed the practical need for a safety-oriented pipeline that links visual accident perception with low-latency roadside warning dissemination.

The experimental evaluation on the traffic video dataset demonstrated that the proposed hybrid RSU framework outperformed the single-modality baselines. In the reported results, the method achieved 95.8 % accuracy, 95.0 % precision, 94.6 % recall, and 94.8 % F1-score, while the average processing time was approximately 40 ms per frame on the considered RSU prototype. These findings indicate that the fusion of spatial and motion information provides a favorable balance between detection quality and real-time execution requirements for intelligent transportation applications.

From an application perspective, the study shows that RSUs can serve not only as communication relays, but also as local edge-intelligence nodes capable of detecting hazardous traffic events and disseminating timely safety alerts to nearby vehicles. This makes the proposed framework especially relevant for intersections, highway segments, ramps, and other safety-critical road locations where rapid accident recognition and warning delivery are essential.

At the same time, the present study has several limitations. First, the current framework was evaluated under a constrained experimental setup and therefore requires further validation under more diverse acquisition conditions and larger deployment scenarios. Second, severe adverse weather conditions, including heavy rain, fog, and snow, may reduce the reliability of visual perception and optical-flow estimation. Third, the current implementation is centered on single-RSU operation and does not yet address large-scale cooperative roadside deployment.

Future research should therefore focus on multi-RSU cooperation, adaptive thresholding under changing traffic conditions, and multimodal sensing through radar, thermal cameras, or other infrastructure-level sensors. Another important direction is privacy-aware large-scale model

updating, including federated-learning-based strategies that can improve deployment scalability without centralizing raw roadside video data [26-28]. In addition, secure large-scale deployment of RSU-based systems may benefit from advances in next-generation vehicular networking, heterogeneous communication management, secure IoT communication, and AI-assisted cyberprotection [28-35]. Overall, the proposed framework provides a practical foundation for integrating computer vision, edge intelligence, and V2X communication in next-generation smart-road infrastructure.

Acknowledgements

The authors have not disclosed any funding.

Data availability

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Author contributions

Danish Ather: conceptualization; methodology; system architecture design; formal analysis; supervision; writing-review and editing. Miraziz Talipov: software implementation, data curation, experimental setup, investigation, visualization, writing-original draft.

Conflict of interest

The authors declare that they have no conflict of interest.

References

[1] J. B. Kenney, "Dedicated short-range communications (DSRC) standards in the United States," *Proceedings of the IEEE*, Vol. 99, No. 7, pp. 1162–1182, Jul. 2011, https://doi.org/10.1109/jproc.2011.2132790

[2] C. Campolo, A. Molinaro, and R. Scopigno, "From Today's VANETs to Tomorrow's planning and the bets for the day after," *Vehicular Communications*, Vol. 2, No. 3, pp. 158–171, Jul. 2015, https://doi.org/10.1016/j.vehcom.2015.06.002

[3] D. Raj, D. Ather, and A. K. Sagar, "Advancing vehicular ad-hoc network solutions in emerging economies: a comparative analysis of V2V protocols through simulation studies," *SN Computer Science*, Vol. 5, No. 8, Nov. 2024, https://doi.org/10.1007/s42979-024-03411-1

[4] M. H. C. Garcia et al., "A tutorial on 5G NR V2X communications," *IEEE Communications Surveys and Tutorials*, Vol. 23, No. 3, pp. 1972–2026, 2021, https://doi.org/10.1109/comst.2021.3057017

[5] "Intelligent transport systems (ITS); vehicular communications; basic set of applications; decentralized environmental notification service (DENM); release 2," ETSI, TS 103 831 V2.2.1, Apr. 2024.

[6] T. Gong, L. Zhu, F. R. Yu, and T. Tang, "Edge intelligence in intelligent transportation systems: a survey," *IEEE Transactions on Intelligent Transportation Systems*, Vol. 24, No. 9, pp. 8919–8944, Sep. 2023, https://doi.org/10.1109/tits.2023.3275741

[7] C. Creß, Z. Bing, and A. C. Knoll, "Intelligent transportation systems using roadside infrastructure: a literature survey," *IEEE Transactions on Intelligent Transportation Systems*, Vol. 25, No. 7, pp. 6309–6327, Jul. 2024, https://doi.org/10.1109/tits.2023.3343434

[8] Y. Ji et al., "Toward autonomous vehicles: a survey on cooperative vehicle-infrastructure system," *IScience*, Vol. 27, No. 5, p. 109751, May 2024, https://doi.org/10.1016/j.isci.2024.109751

[9] G. Cui, W. Zhang, Y. Xiao, L. Yao, and Z. Fang, "Cooperative perception technology of autonomous driving in the Internet of Vehicles environment: a review," *Sensors*, Vol. 22, No. 15, p. 5535, Jul. 2022, https://doi.org/10.3390/s22155535

[10] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779–788, 2016, https://doi.org/10.1109/cvpr.2016.91

[11] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Image Analysis*, pp. 363–370, 2003, https://doi.org/10.1007/3-540-45103-x_50

[12] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8934–8943, 2018, https://doi.org/10.1109/cvpr.2018.00931

[13] J. Fang, J. Qiao, J. Xue, and Z. Li, "Vision-based traffic accident detection and anticipation: a survey," *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 34, No. 4, pp. 1983–1999, Apr. 2024, https://doi.org/10.1109/tcsvt.2023.3307655

[14] T. Huang et al., "Vehicle-to-everything cooperative perception for autonomous driving," *ArXiv*, 2023, https://doi.org/10.48550/arxiv.2310.03525

[15] S. Masi, P. Xu, P. Bonnifait, and S.-S. Ieng, "Augmented perception with cooperative roadside vision systems for autonomous driving in complex scenarios," in *IEEE International Intelligent Transportation Systems Conference (ITSC)*, pp. 1140–1146, 2021, https://doi.org/10.1109/itsc48978.2021.9564833

[16] R. Zhang et al., "Evaluating roadside perception for autonomous vehicles: insights from field testing," *ArXiv*, 2024, https://doi.org/10.48550/arxiv.2401.12392

[17] Y. Mo, R. Vijay, R. Rufus, N. Boer, J. Kim, and M. Yu, "Enhanced perception for autonomous vehicles at obstructed intersections: an implementation of vehicle to infrastructure (V2I) collaboration," *Sensors*, Vol. 24, No. 3, p. 936, Jan. 2024, https://doi.org/10.3390/s24030936

[18] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: the KITTI dataset," *The International Journal of Robotics Research*, Vol. 32, No. 11, pp. 1231–1237, Sep. 2013, https://doi.org/10.1177/0278364913491297

[19] H. Yu et al., "DAIR-V2X: a large-scale dataset for vehicle-infrastructure cooperative 3D object detection," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21329–21338, 2022, https://doi.org/10.1109/cvpr52688.2022.02067

[20] H. Yu et al., "V2X-Seq: a large-scale sequential dataset for vehicle-infrastructure cooperative perception and forecasting," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5486–5495, 2023, https://doi.org/10.1109/cvpr52729.2023.00531

[21] W. Zimmer, G. A. Wardana, S. Sritharan, X. Zhou, R. Song, and A. C. Knoll, "TUMTraf V2X cooperative perception dataset," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22668–22677, 2024, https://doi.org/10.1109/cvpr52733.2024.02139

[22] H. Xiang et al., "V2X-Real: a large-scale dataset for vehicle-to-everything cooperative perception," in *Computer Vision – ECCV*, pp. 455–470, 2024, https://doi.org/10.1007/978-3-031-72943-0_26

[23] M. Talipov, "Computational modeling and analysis of mechanical power consumption in train assemblers' work," in *International Conference on Applied Innovations in IT*, Vol. 13, No. 2, pp. 419–426, Jun. 2025, https://doi.org/10.25673/120513

[24] S. S. Sulaymonov, S. K. Abdazimov, and X. G. Azimov, "Modern approach to environmental emergency warning at high facilities," *E3S Web of Conferences*, Vol. 477, p. 00103, 2024, https://doi.org/10.1051/e3sconf/202447700103

[25] S. S. Sulaymanov, K. M. Nurmatov, and N. B. Chetkova, "Strengthening the capacity of infrastructure facilities of the Angren-Pap railway line to reduce the seismic risk," in *AIP Conference Proceedings*, Vol. 2943, No. 1, p. 030056, 2023, https://doi.org/10.1063/5.0141972

[26] K. Bonawitz et al., "Towards federated learning at scale: system design," in *Proceedings of Machine Learning and Systems*, Vol. 1, pp. 374–388, 2019.

[27] A. Bhardwaj, A. Sharma, D. Raj, D. Ather, A. K. Sagar, and V. Jain, "Dynamic and scalable privacy-preserving group data sharing in secure cloud computing," in *Advances in Information Security, Privacy, and Ethics*, Vol. 1, Hershey, PA, USA: IGI Global, 2024, pp. 89–122, https://doi.org/10.4018/979-8-3693-9225-6.ch004

[28] D. Ather, H. Kaur, M. Rakhra, R. Kler, G. Aggarwal, and K. Jairath, "Data privacy in edge computing: securing solutions for distributed systems," in *International Conference on Networks and Cryptology (NETCRYPT)*, pp. 1747–1752, 2025, https://doi.org/10.1109/netcrypt65877.2025.11102306

[29] D. Ather, R. Kler, Z. Tanveer Baig, G. Prakash Babu, A. Rastogi, and N. Ahmed, "6G networks," in *Network Security and Data Privacy in 6G Communication*, New York: CRC Press, 2025, pp. 234–253, https://doi.org/10.1201/9781003583127-12

[30] D. Raj, A. K. Sagar, and D. Ather, "Optimized channel allocation algorithm for performance enhancement in heterogeneous networks," *MATRIX Academic International Online Journal of Engineering and Technology*, Vol. 8, No. 1, pp. 1–12, 2025, https://doi.org/10.21276/matrix.2025.8.1.3

[31] D. Raj, A. K. Sagar, and D. Ather, "Optimizing vertical handover using multi-criteria decision making in heterogeneous networks," *Information Technology and Management*, Vol. 17, No. Special Issue, pp. 63–86, 2025.

[32] M. Rakhra, B. Kaur, G. Aggarwal, D. Ather, R. Kler, and K. Jairath, "The zero trust paradigm: revolutionizing network security," in *International Conference on Networks and Cryptology (NETCRYPT)*, pp. 1714–1719, 2025, https://doi.org/10.1109/netcrypt65877.2025.11102576

[33] B. Kaur, M. Rakhra, R. Kler, G. Aggarwal, D. Ather, and K. Jairath, "Secure IoT communication protocols: a comprehensive framework for smart devices," in *International Conference on Networks and Cryptology (NETCRYPT)*, pp. 390–396, 2025, https://doi.org/10.1109/netcrypt65877.2025.11102768

[34] D. Ather, A. Singh, M. Rakhra, R. Kler, G. Aggarwal, and K. Jairath, "AI-driven threat intelligence for proactive cybersecurity in enterprise networks," in *International Conference on Networks and Cryptology (NETCRYPT)*, pp. 1720–1725, 2025, https://doi.org/10.1109/netcrypt65877.2025.11102400

[35] M. Rakhra, R. Kanday, G. Aggarwal, D. Ather, R. Kler, and K. Jairath, "Quantum cryptography: enhancing secure communication in the era of quantum computing," in *International Conference on Networks and Cryptology (NETCRYPT)*, pp. 1395–1398, 2025, https://doi.org/10.1109/netcrypt65877.2025.11102170

Appendix

A1. Model compression and optimization

To further decrease both the time and memory footprint inference, standard model-compression methods may be used:

To minimize still further the time of inference and memory footprint, the typical model-compression methods can be used: 8-bit quantization of network parameters and activations; Layer fusion and operator reordering for better cache utilization.

Nearly these techniques can be implemented during a post-training phase which does not influence the performance of detecting significantly.

A2. Handling missing or corrupted frames

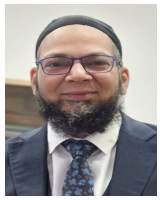
In actual implementations lost or corrupted frames can occur as a result of network or equipment problems. One easy trick is to continue the valid optical flow and CNN features frame by frame over a restricted sequence of time frames or to temporally adopt spatial-only detection until motion data is again available.

A3. Additional numerical results

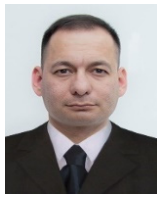
Table A1 presents a sensitivity analysis model of the proposed framework to the averaging window size of the time period. K in Eq. (13).

Table A1. Illustrative effect of temporal averaging window K on performance metrics

Window size K	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
$K = 1$ (no averaging)	94.0	93.0	92.4	92.7
$K = 3$	95.2	94.5	93.8	94.1
$K = 5$	95.8	95.0	94.6	94.8
$K = 7$	95.6	95.3	94.0	94.6



Danish Ather is a researcher at Amity University, Noida, India. His work focuses on Intelligent Transportation Systems (ITS), the Internet of Things (IoT), and AI-based decision support. He has contributed to numerous projects involving smart cities, real-time sensing, and edge computing. His recent research investigates AI-driven perception and vehicular communication to enhance the safety and efficiency of road networks.



Miraziz Talipov, Ph.D. Doctor of Philosophy in Technical Sciences, specializing in Occupational Health and Human Safety. Currently a researcher at Tashkent State Transport University, Dr. Talipov focuses on the intersection of industrial safety and Intelligent Transportation Systems (ITS). His expertise covers risk assessment, ergonomics, and the deployment of V2X communication, RSU strategies, and advanced sensing technologies to modernize road infrastructure and enhance transportation safety.